Distr.
GENERAL

Working Paper No.7
19 March 2008

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)

STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2007)
(Luxembourg, 7-9 May 2008)

Topic (ii): Statistical information systems architecture

METADATA ARCHITECTURE AT STATISTICS CANADA

Invited Paper

Prepared by Karen Doherty, Statistics Canada, Canada

I. INTRODUCTION

1. Statistics Canada (STC) maintains a large collection of data offerings. To maximize the understandability of these data files, STC has put considerable effort into the development of metadata for each data offering. At first the emphasis was to describe the nature of the surveys and administrative sources from which the data was derived. In recent years the focus has shifted towards the detailed description of the data variables that make up a data offering and the definition and meaning of the code sets associated with these variables.

2. In parallel to this initiative Statistics Canada embarked on the development of a flexible architecture for analytical data warehouses. This initiative proved to be extremely successful and warehouses have proliferated across the Agency, not only for analytical purposes, but also as tools to provide up-to-date information to managers in both traditional management roles and operations. As analysts became familiar with the warehousing tools, they starting asking for real-time access to the metadata describing the data they were manipulating.

3. To make metadata accessible to applications within STC the IT team responsible for data warehousing began investigating easy-to-use solutions to this problem. Over time this work has led to a vision for the support and use of metadata across the Agency. Although the architecture is not yet fully implemented the intent is to have this architecture adopted as an Agency standard. This report explains the components of the architecture and the benefits that are being achieved.

II. THE BUILDING BLOCKS

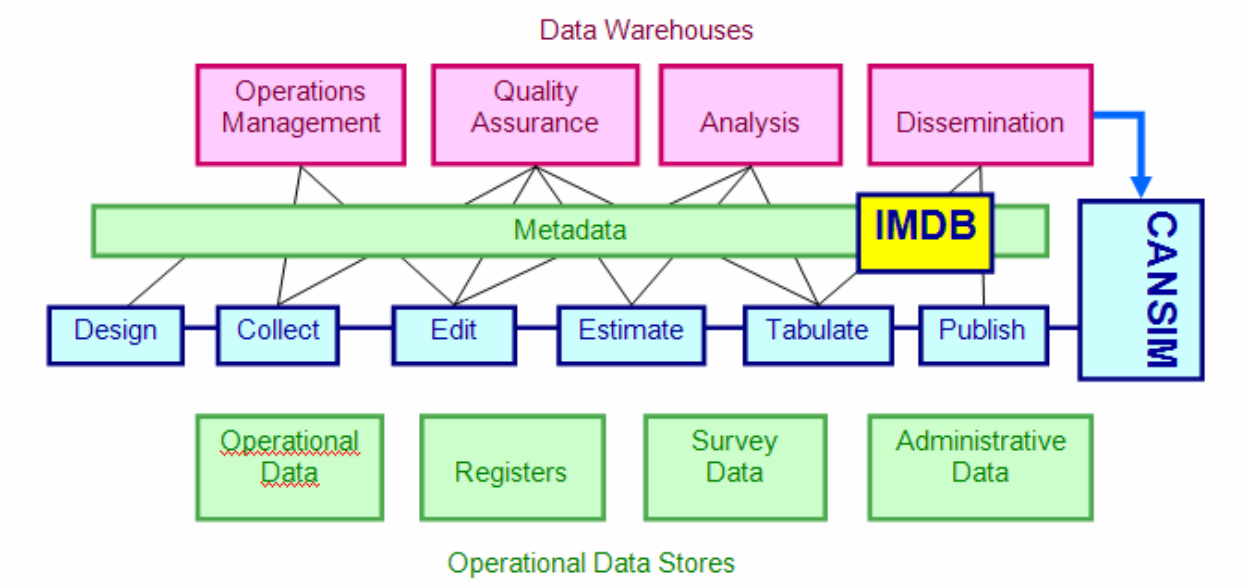
4. Between 1998 and 2006 a number of systems and system components related to the management of metadata were developed to address specific business needs within the Agency. Although these initiatives did not necessary lead to system duplication, the result was a confusion of different pieces with no coherent plan or vision for how metadata systems should be managed.

5. This section of the report describes the various services and systems which were developed more or less independently but which all touch directly or indirectly on metadata issues.

A. The Integrated Metadata Bases (IMDB)

6. The Integrated Metadata Base (IMDB) is a corporate repository of information on each of Statistics Canada's nearly 400 active surveys (and a roughly equal number of discontinued surveys). These surveys are the Agency's core activities and the IMDB is the principal mechanism by which they are documented, providing a key information resource for corporate knowledge management and for data users.

7. Metadata can support at least three broad functions within a statistical agency: data dissemination; data collection and production, and management of the statistical system. The original vision for the IMDB was to provide metadata as needed at any phase, be it collection, processing or dissemination. When the IMDB was first conceived the greatest need was the support of the dissemination function, in particular the provision of microdata to describe the contents of CANSIM. The team therefore chose to implement the funcitonality needed to provide users with the information required to interpret the statistical data we disseminate. As a result, the original content of the database was largely dictated by the requirements of STC's *Policy on Informing Users of Data Quality and Methodology*. However, the team has never lost sight of the long term vision which is represented by the following diagram.



8. The database is resident on a central server. For the initial load of this database, metadata were collected from a variety of pre-existing metadata stores, reformated, validated and loaded into the new metadata base. The database is kept up to date by Standards Division staff, who obtain the necessary information from specialists in subject-matter divisions and input this information through a system deployed over the departmental Intranet. Formatted HTML pages of the IMDB content are made available on the Statistics Canada web site. They can be accessed through hyperlinks from CANSIM (STC's data dissemination tool), from the online catalogue or from statistical tables on the web site. The pages can also be accessed directly through a thematic or alphabetic index in the *Definitions, data sources and methods* module of the STC web site.

9. The content of the metadata base has been selected to suit its primary purpose, which is to provide users with the information required for the interpretation of statistical data we disseminate. The IMDB data model can be thought of as consisting of two regions. The first, the statistical information region, defines all of the entities and the relationships among them required for describing Statistics Canada's statistical programs, their content and their methods. The second, the data definitional region, defines the concepts and classifications for survey variables (referred to as data elements in IMDB terminology).

10. The IMDB's data model is based on the ISO 11179.

11. IMDB entities are referred to as an administered item. As administered items, they all share an identical set of characteristics for the stewardship of the component. This stewardship region includes such information as theme, keyword, time frames, contact, organization and documentation.

B. ISO 11179

12. The IMDB was originally designed to be a repository, or Metadata Registry, of the metadata describing our surveys and survey outputs. However, since metadata by definition can be collected on any group of objects, STC has begun to expand the scope of the IMDB to be a repository of objects other than our data outputs. In particular, we have begun to view structured information describing our IT systems and our enterprise architecture as metadata.

13. The ability to consider expanding the scope of the information housed in the IMDB is possible because the IMDB conforms to a strongly defined metadata standard, ISO 11179. Although this standard was designed specifically to support metadata on data, the standard can easily be applied to any group of objects for which metadata is being collected.

14. There are four main concepts used to describe a "data element" or object:

- **Object class**: the object class refers to the type of object being described, an entity such as a person, an establishment or a household.
- **Property**: an object can have one or more properties. Each property is an attribute of the object, for instance, a "person" can have several attributes or properties such as sex, age, height, etc.
- **Conceptual Domain**: each property could have several possible value meanings. Meanings are generally articulated in real language and cover all the possible values that are meaningful for that property. For instance, the sex (property) of a person (class) can have two value meanings, male or female.
- **Value Domain**: in general computer systems do not store value meanings but rather value codes. Value codes are the permissible codes used to represent the value meanings. For instance a particular implementation of the sex property in a data file may represent the value Male by a 1, and the value Female by a 2.

15. The example above refers to metadata as it applies to statistical data. Here the object "Person" has a property "Sex" which has two possible values meanings "Male" and "Female" represented respectively by the value codes "1" and "2". The standard could just as well be used to describe metadata about our Enterprise Architecture. For example, an object "Application" could have a property called "Development Language" with several value meanings such as "MS VB.net", "SAS", "Java", etc. with corresponding value codes.

16. ISO 11179 allows STC to represent all the data elements in its data output using a consistent, structured set of concepts and definitions. Although the standard can be difficult to master, it allows metadata to be provided to end-user applications independently of the platform or products used to implement these end-user applications. For example, the IMDB can feed CANSIM, STC's data series repository used for dissemination purposes, any warehouse built using the STC Data Warehouse Framework and STCwiki which provides users and IT professionals with a collaborative environment in which to display and discuss metadata. By marrying metadata information from the IMDB with the associated data file, it is also possible to automatically create output files that conform to other data reporting formats such as DDI and SDMX.

C. Data Warehouses

17. In 2001, the Agency began developing a data warehouse to provide analysts working with Canadian System of National Accounts (SNA) data with a conceptually integrated framework of statistics and analysis for studying the state and behaviour of the Canadian economy. The warehouse needed to be able to resolve issues related to data inputs and suppliers and provide analysts with a mechanism to review and improve of data outputs and products.

18. This initiative along with subsequent warehousing initiatives within the Agency resulted in the development of STC's Data Warehouse Framework which was adopted as the corporate standard in 2005. STC analytical warehouses allow users to compare statistics in terms of ratios, proportions, growth rates, by region and in chronological series. They are particularly helpful in enhancing data coherency and in permitting comparisons of data taken from different sources, while applying classifications, definitions and concepts for any aggregate level. Access to metadata is provided, as is information on how the data was established, concepts and definitions, classifications and concordances and best practices with respect to processing or analytical procedures.

19. STC's analytical data warehouses contain vast amounts of survey microdata and aggregated data from a variety of sources. Analysts using these warehouses quickly discovered that they could progress much faster if they had direct access to the metadata describing the information they were working with. This provided STC with an opportunity to move a step closer to the original vision for the IMDB, the use of the IMDB across all phases of the survey life cycle.

D. EzWeb

20. Developed and maintained by Statistics Canada, EzWeb is a product used for intranet web site construction and management. Using a simple "what you see is what you get" (WYSIWYG) approach, EzWeb provides users who have no web skills with a user-friendly environment in which to create web pages that are compliant with the Canadian Government's Government On Line (GOL) standard, including the need to make pages available in both official languages.

21. Microsoft Office Web Components (OWC) Pivot Table is incorporated into EzWeb and provides the OLAP functionality needed to view data in warehouse cubes. EzWeb also has the ability to navigate from one OLAP report to another, thus providing users with easy and intuitive information exploration and discovery through a single interface.

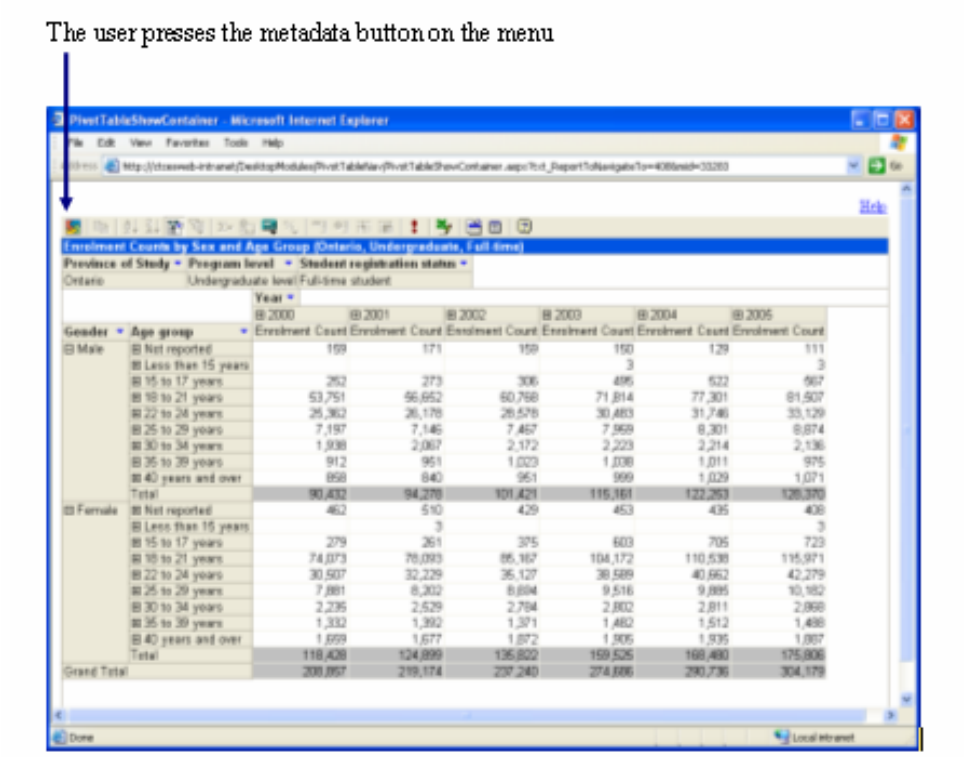
D. STCwiki

22. One of the challenges faced by the warehouse architects was how to provide warehouse users with access to metadata in a way that allows them to navigate quickly through the rich metadata environment and to submit updates to Standards Division related to metadata that is incomplete or inaccurate. The DW Centre resolved this issue by leveraging a wiki application to act as a link between the Agency's Integrated Meta Data Base (IMDB) and the Data Warehouse Framework.

23. A wiki is a tool for collaborative authoring and knowledge sharing. The most well-known wiki application is Wikipedia, the online collaborative encyclopedia which contains over 1.7 million articles. There are a number of wiki engines available in the Open Source domain – STC chose MediaWiki (used by Wikipedia) to build STCwiki.

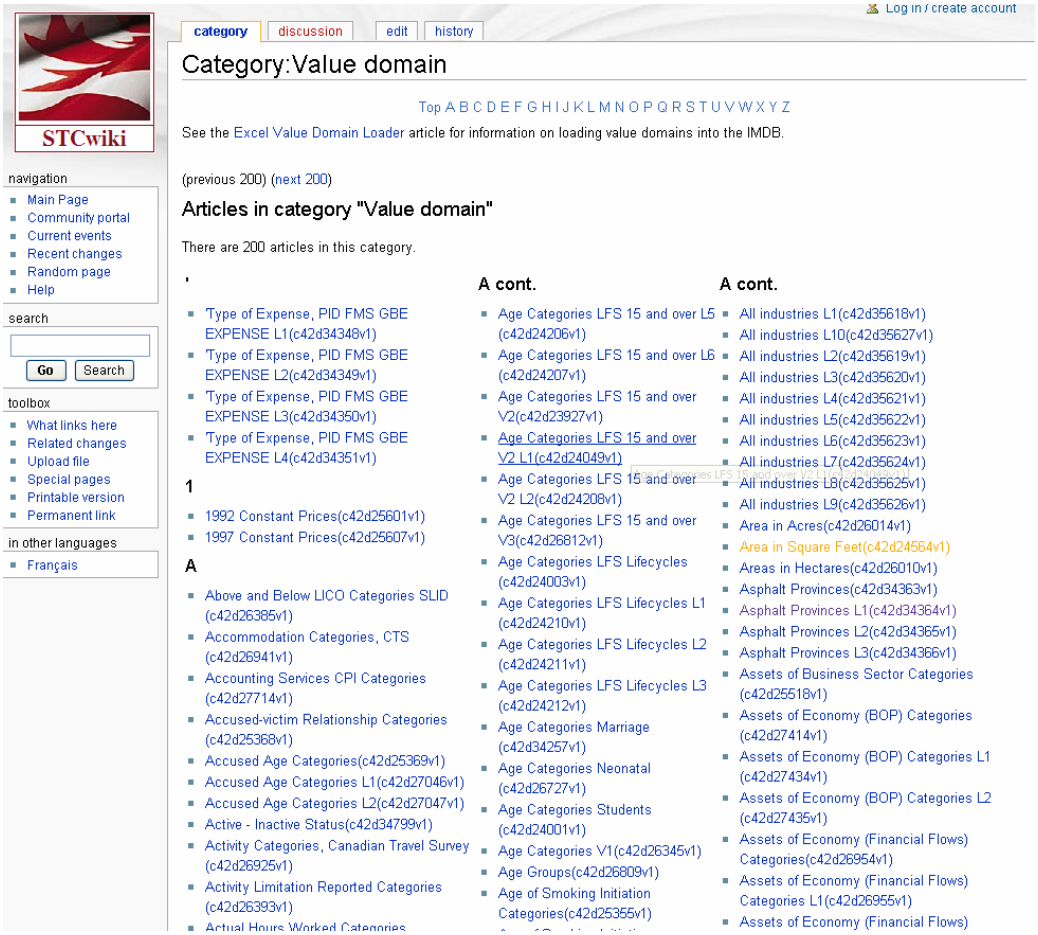
III. VIEWING METADATA VIA THE STCWIKI

24. The introduction of the STCwiki has made it much easier for users to view metadata either while working in a warehouse environment or in a standalone mode. From a warehouse the user has direct access to the metadata via a menu button on the view screen.



25. The menu button accesses the STCwiki and displays the specific metadata they wish to consult. They can also broaden their search as needed to retrieve metadata about other variables or surveys.

26. The STCwiki also gives users access to a wealth of information via traditional menus and search functions.



IV. THE METADATA ARCHITECTURE

27. By 2007 it was apparent that it was time for the Agency to start to leverage our investment in the Integrated Metadata Base system to not only supply metadata to our dissemination systems (CANSIM) and analytical warehouses but also to address other capabilities such as the support for SDMX compliant data files and the documentation of metadata about IT systems and architecture.

28. The realisation that this was a good opportunity to expand the role of IMDB resulted in a collaboration between the teams responsible for data warehousing (the Data Warehouse Centre) and architecture (the Centre for Architecture) to rearticulate the long term vision for the management of metadata systems at STC.

A. Objectives

29. The objectives of the various initiatives were:

(i) Leverage the investment in the IMDB's ISO 11179 compliant metadata architecture to expand beyond the scope of Statistics Canada's statistical data products, in particular to start documenting metadata about STC's IT systems and enterprise architecture.

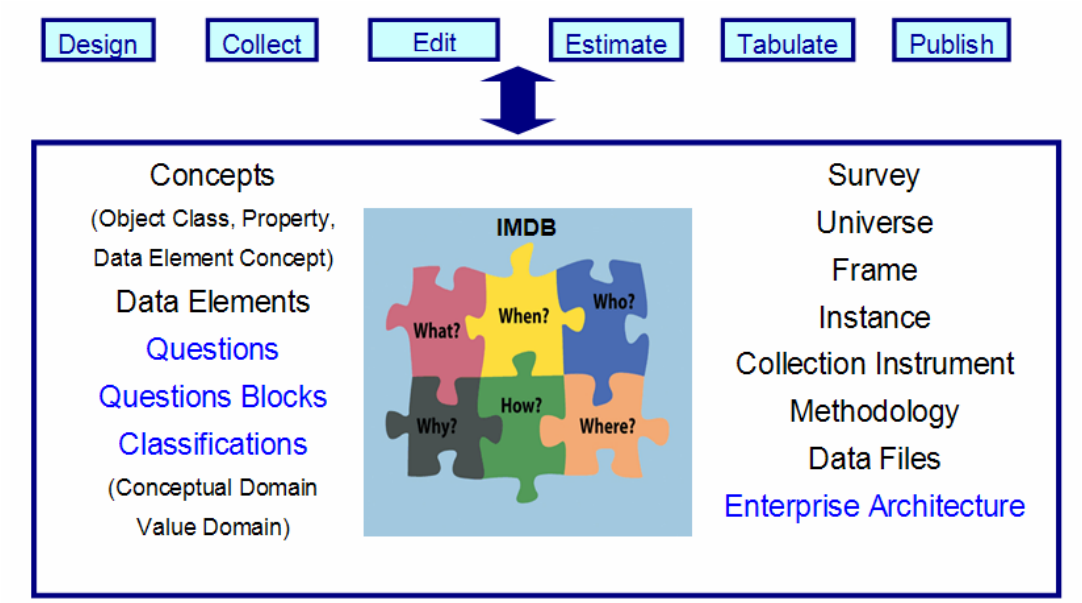
(ii) Develop a user-friendly interface for viewing the metadata directly or within a data warehouse.

(iii) Improve the maintenance and support of coding activities, in particular those using standard classification systems such as NAICS.

(iv) Develop a user-friendly mechanism for analysts and users to suggest updates to the metadata and manage the change control process.

B. Expansion of the Role of the IMDB

30. Once the idea took hold that the IMDB could be used to collect and feed metadata to users at any step in the survey life cycle the vision for the role of the IMDB in the Agency's Enterprise Architecture expanded enormously.



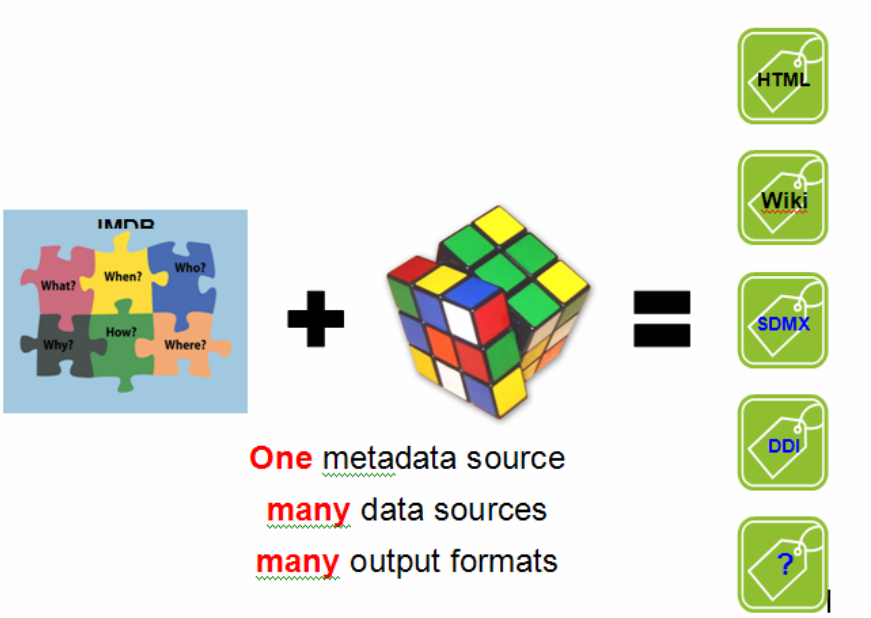
31. Firstly, an expanded view of the role in the IMDB with respect to the support of standard Classifications expanded from the limited view of simply supporting the requirements of CANSIM to the idea of publishing classification information directly to the STC Web Site from the IMDB and from there to supporting generic coding applications.

32. Next came the notion of maintaining standard questions and question blocks to support the creation of questionnaires. Although there are many challenges associated with this step, the Agency has begun the process of articulating requirements and possibilities in this area.

33. And finally, the ability to display descriptions contained in the IMDB about our business and surveys via STCwiki, led to a project to expand on this theme by using the IMDB as the repository of the documentation of our Enterprise Architecture.

C. Producing Output Files in Various Formats

34. The IMDB increasingly contains very detailed information about our surveys including the variables, classifications and codes set that describe our publishable outputs. More and more of the actual data is now being housed in our data warehouses. Analysts can produce output files within the warehouse environment and, via STCwiki verify and suggest updates to, the metadata that describes these files. Since the IMDB information is maintained using a comprehensive standard (ISO 11197) and the warehouse data is arranged to work in tandem with the metadata the next step is to produce output files which combine the metadata and the data and render the result in a format that can be used by other applications. In particular, the IMDB development team is now working on the high level design of a module that would render such a file in SDMX format.

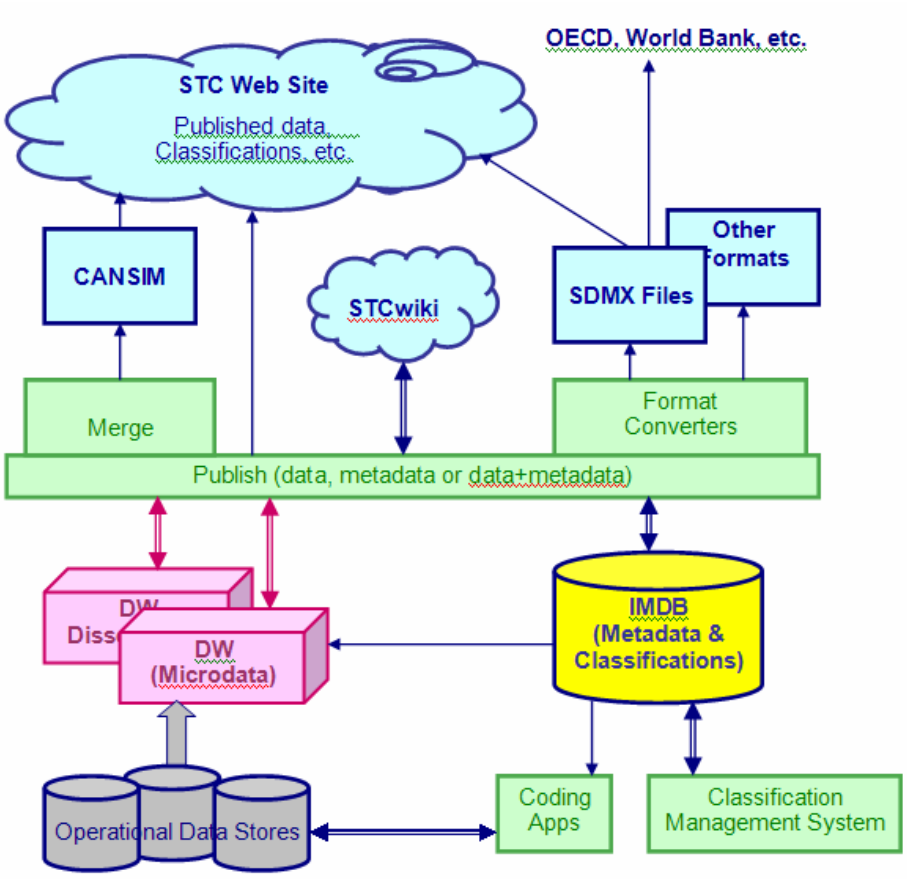


D. Architecture Vision – Phase 1

35. The following diagram describes the components of the IMDB architecture vision that are already in place, under development or planned for implementation within the next two years.

36. In yellow is the IMDB repository and supporting user interfaces. The Classification Management System (bottom in green) is the application which will maintain the standard classifications and their associated reference files, etc. The IMDB will feed classification data to generic coding applications used to code data being processed in the operational data stores (bottom left in grey). The data in the operational data stores feed data into the data warehouses. IMDB can also feed metadata directly into the data warehouses.

37. Applications (in the middle in green) retrieve data from the warehouses and from the IMDB to produce a large variety of outputs. The data can be combined to feed CANSIM which provides public access to published data via the STC Web Site. Classification tables from the IMDB can also be published on the STC Web Site. Output files contain data and metadata in SDMX or other formats can be created for publishing on the web site or for distribution to other organizations. Finally, metadata can be posted on STCwiki allowing users to view the information and submit changes to the definitions and descriptions.



38. Although much of the target architecture has already been implemented many of the components are still under development or are planned for development over the next two years. Subsequent phases will look at expanding the use of the metadata in the IMDB, in particular by incorporating questions and question blocks as well as code sets into the collection and processing steps in the survey life cycle.

39. Work already completed:

- The IMDB repository and supporting software
- The Data Warehouse Framework
- STCwiki

40. Work in progress:

- The Classification Management System
- Generic coding applications
- Loading the Enterprise Architecture information into the IMDB

41. Work still in the planning stages:

- Publishing classification data automatically from the IMDB
- The SDMX formatter
- The Data Warehouse CANSIM loader

V. CONCLUSION

42. Over the past seven years, the view of the IMDB's role has evolved from being simply a mechanism to feed metadata to CANSIM for the publication of output tables, to being an integral part of the applications that manage the survey life cycle. This change can be in large part attributed to popularity of the data warehouse initiatives over the past five years and the associated evolution of the data warehouse services.

43. There is still a lot of work to be done to fully realize the potential of the Metadata vision and we will undoubtedly face many challenges along the way but the early indicators are positive and we are confident that the initiatives will both improve applications available to STC employees and deliver savings through increased automation and reduced duplication of processes and systems.